203 Wadleigh's Falls Road
Lee, NH 03861
January 26, 2009

Mr. Wayne Ives, P.G.
NHDES
P.O. Box 95
29 Hazen Drive
Concord, NH 03302-0095

RE: Draft Report on Lamprey River Protected Instream Flow Study
Dear Mr. Ives,

I am writing in response to your request for comments on the above study made during the public presentation at the Lee Safety Complex. As a mathematician I have been asked by the LRAC to provide an assessment and overview of the modeling portion of the report conducted by Rushing Rivers Institute. I have been working through the report and have had a detailed conversation with Piotr Parasiewicz concerning the methods and procedures used in the various components of the study. As a result, I have some comments which I believe should be addressed in the final report. In order to describe my concerns I will begin by summarizing the overall procedure as I now understand it so that if there is a misunderstanding there it can be easily addressed.

As I understand the modeling effort (for each fish species or group):

1. Data was collected from a number of river studies which include a variety of physical and biological variables considered as independent and fish counts as dependent variables.
2. A randomly selected validation set of 20% of the data was extracted and a model estimated using the remaining 80% of the data. The variables to be included in the model were selected based on the Akaike Information Criterion (AIC) and the selected model was tested for its ability to correctly identify fish observations in the validation set. The selected independent variables were recorded.
3. Step 2 was repeated 20 times generating 20 sets of independent variables and 20 "success rates". A final model was then constructed using the AIC procedure with possible independent variables restricted to those which had occurred at least 2 times among the 20 variable sets generated by the preliminary models. The final model was accepted if its "success rate" was considered sufficiently close to the average of the preliminary 20 "success rates".
4. The final model was used to estimate habitat suitability in each of the Lamprey River study segments using the appropriate local values of the model's independent variables.
5. The final model was used to estimate baseline habitat suitability in each of the Lamprey River study segments using the appropriate local values of the model's independent variables modified to represent the modeler's best estimate of the baseline natural conditions.
6. The Protected Instream Flow recommendations are based on the baseline habitat suitability curves from Step 5 and other considerations.

I begin my comments by congratulating the modelers on their efforts in Steps 1 through 3. Seldom does one see such a careful model validation process. Their work is exceptional in that regard as is shown by the high success rates their models achieve. However, I do have comments which I think are important and should be addressed.

1. In general, when using the AIC procedure with a great many variables the values to be minimized can be vary close together with many model values, each model with different

independent variables, differing by only small and, probably, insignificant amounts. The AIC process is intended for comparing model constructions and does not, necessarily, identify the most understandable and explanatory model. This becomes very apparent when one tries to understand what the presented models are actually indicating. For example, the first "presence" model summary of the report shows two coefficients of -234 and 20 while the remaining variable coefficients are all less than 3 in absolute value. The variable with the 20 coefficient becomes the largest in the "abundance" model but changes sign now indicating a negative influence on abundance rather than its positive influence on presence. These differences may be justified but from the information presented in the report that can not be determined and so may be a (possibly misleading) statistical artifact.

2. The standard way of addressing the problem just described is by providing an analysis of variance table showing the percent of variance explained by each of the explanatory variables and/or estimated standard deviations of the coefficients. With that information one can assess the importance of each variable in understanding the system under study. In this situation one does not even know how many of the 20 preliminary variable sets contained each of the final variables. There is a comment in the report that some of the "large coefficients" are statistical artifacts arising from single observations. If that is indeed the case, those coefficients should not be included in the final report as they are very misleading. How large are "large coefficients"? I think a standard logistic regression modeling effort should be conducted using the variables identified and used in each "final" model. With the results of that analysis one would have a much better idea of the important explanatory variables and the variation in the estimated coefficients for use in the following parts of the study.
3. When a model contains many variables which explain very small and statistically insignificant amounts of variance the coefficients of the truly important explanatory variables can be considerably different when the extraneous variables are omitted. In the models reported here we have no way of judging the validity of the estimated coefficients. They may, or may not, deserve the importance placed on them in the habitat suitability estimates for which they are used. Without some estimate of significance we can not be confident in the results.
4. These concerns are important as the estimated coefficients are used to assess habitat suitability for the Lamprey River sections and, in turn, are used in defining the PISF. Unless confidence can be placed on the validity of the model coefficients there can be considerable doubt as to the value of the final product.

In summary, it is my opinion that the models were developed in a semi-automatic way using the AIC process which requires little supervision. While that process is an excellent way of developing preliminary models, it is, in my opinion, not automatically acceptable in generating models which succinctly summarize and provide understanding of the process under study. I believe it is the latter that is needed for this study as its final conclusions, the PISF, seem to rely strongly on the model results.

It is, of course, possible that I do not understand all aspects of this study. If so, or in any case, I would be pleased to discuss the issues raised here with you or others.

Sincerely,

Dr. Loren D. Meeker
Professor Emeritus of Mathematics
UNH